BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25438
doi: 10.3897/biss.2.25438

OPEN ACCESS

Conference Abstract

Vocabularies of Values: Tackling the Heterogeneity Problem

Paula F Zermoglio [‡]

[‡] Instituto de Ecología, Genética y Evolución de Buenos Aires (IEGEBA-CONICET), University of Buenos Aires, , Argentina

Corresponding author: Paula F Zermoglio (pzermoglio@gmail.com)

Received: 02 Apr 2018 | Published: 18 May 2018

Citation: Zermoglio P (2018) Vocabularies of Values: Tackling the Heterogeneity Problem. Biodiversity Information Science and Standards 2: e25438. https://doi.org/10.3897/biss.2.25438

Abstract

In the process of sharing information, it is of highest importance that we utilize common codes and signifiers, so that communication is effective. This process presents a series of complexities that are related to capturing and transmitting the meaning of the information despite homonymy, polysemy and synonymy. Biodiversity data sharing is not exempt from these challenges and understanding the meaning often requires expert knowledge. For communication to be effective, and therefore for data to be of maximal re-use, we need common vocabularies that unequivocally refer us to the same concepts.

The community has agreed upon some vocabularies to structure shared information, i.e., biodiversity data standards such as the Darwin Core standard (Wieczorek et al. 2012). The bterms in Darwin Core can be thought of as the names of the columns in a spreadsheet. For example, there are terms such as genus, stateProvince, sex, etc. This allows us to capture and share information which we agree belongs under one of those terms. However, we have not yet reached an agreement on how to express the permitted values under all those terms, that is, vocabularies of values. As a simple example, we agree that if we have a record of an organism that is a *female*, we will share the fact that it is a female under the “sex” term, but we could represent *female* with the values “female”, “fem.”, “f.”, and other possible abbreviation and language variants. Other more complex examples, bound to expert knowledge, include biological taxonomies and how we name distinct species and species concepts.

While many vocabularies exist in the community, we currently do not possess a full suite of vocabularies of values that apply uniformly across the biodiversity data community and there is no single repository to explore the available resources. While some of the available vocabularies are discipline-specific, many that could be applied more broadly remain independent and scattered. Additionally, similar lists of terms that refer to the same concepts can be found in different languages, but disconnected from one another.

The lack of or non-adherence to vocabularies of values constitutes a data quality issue, as the heterogeneity in the data renders data less discoverable and difficult to use. Capturing information in myriad ways risks being incomplete and inaccurate in our transmission of information. If we cannot be certain that a particular value unambiguously refers to a particular concept, we cannot assert that a record containing that value could reliably be used for a particular purpose. In this context, the construction and use of vocabularies of values, including the explicit declaration of usage, is a data quality issue.

From the TDWG Data Quality Interest Group we have begun to tackle this problem, with the aim of creating a suitable environment for thought and development of vocabularies of values. Accordingly, a new task group has been constituted, whose main goals are to:

1. prepare a scoping document in which we will determine the types of vocabularies needed (including multi-lingual approaches) and the strategy for organizing the construction and/or management of new/existing vocabularies;
2. develop a common repository to store vocabularies and/or link to existing ones;
3. develop best practices for building TDWG vocabularies; and
4. develop an exemplary vocabulary following the standard format.

This will provide the community with a framework to work on and build upon vocabularies of values in a way that would allow better understanding and maximal interoperability.

Keywords

vocabularies of values, data quality, heterogeneity

Presenting author

Paula F Zermoglio

References

- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1): e29715. https://doi.org/10.1371/journal.pone.0029715